DOCUMENT RESUME

ED 411 254 TM 027 187

AUTHOR Marsh, Herbert W.; Hau, Kit-Tai; Chung, Choi-Man; Siu, Teresa L. P.

TITLE Students' Evaluations of University Teaching: Chinese Version of the Students' Evaluations of Educational Quality (SEEQ) Instrument.

PUB DATE 1996-07-22
NOTE 40p.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Chinese; College Faculty; *College Students; Factor Structure; Foreign Countries; Higher Education; Intellectual Disciplines; *Student Evaluation of Teacher Performance; *Teacher Characteristics; Test Items; *Test Use; Test Validity
IDENTIFIERS Students Evaluation of Educational Quality

ABSTRACT

Chinese University of Hong Kong students (n=844) each rated a "good" teacher and a "poor" teacher using a Chinese translation of the Students' Evaluations of Educational Quality (SEEQ) instrument. Good teachers were rated more favorably than poor teachers on all SEEQ scales, all SEEQ items were judged to be most important by at least some students, and SEEQ items (except, perhaps, feedback on examinations) were seen as appropriate by most students. Relations with background/demographic variables were similar to those reported in North American studies. Multigroup confirmatory factor analysis supported the invariance of SEEQ factor structure across three discipline groups and across ratings of good and poor teachers. In the best model, factor loadings were invariant across all six groups, while factor variances and covariances were invariant across the three discipline groups. The results support the use of the SEEQ in this Chinese setting as well as the generalizability of North American research findings. An appendix presents paraphrased versions of SEEQ items. (Contains 4 tables and 56 references.) (Author/SLD)

Running Head: Students' Evaluations

Students' Evaluations of University Teaching:

Chinese Version of the Students' Evaluations of Educational Quality (SEEQ) Instrument

Herbert W. Marsh

University of Western Sydney at Macarthur, Australia

Kit-Tai Hau, Choi-Man Chung & Teresa L. P. Siu

Chinese University of Hong Kong

22 July, 1996

Abstract

Chinese University of Hong Kong students (N=844) each rated a "good" teacher and a "poor" teacher using a Chinese translation of the Students' Evaluations of Educational Quality (SEEQ) instrument. Good teachers were rated more favorably than poor teachers on all SEEQ scales, all SEEQ items were judged to be most important by at least some students, and SEEQ items (except, perhaps, feedback on examinations) were seen as appropriate by most students. Relations with background/demographic variables were similar to those reported in North American studies. Multi-group confirmatory factor analysis supported the invariance of SEEQ factor structure across three discipline groups and across ratings of good and poor teachers. In the best model, factor loadings were invariant across all six groups, whereas factor variances and covariances were invariant across the three discipline groups. The results support the use of SEEQ in this Chinese setting and the generality of North American research findings.

Students' Evaluations of University Teaching: Chinese Version of the Students' Evaluations of Educational Quality (SEEQ) Instrument

Students' evaluations of teaching (SET) effectiveness are becoming more widely considered for purposes such as feedback to lecturers that may lead to the improvement of teaching, student course selection, personnel decisions, and research on teaching. Particularly in North American universities, SETs are collected almost universally and are the basis of most previous research. This research shows that the ratings are multidimensional (e.g., a teacher may be enthusiastic but lack organization), reliable, stable, reasonably valid against a variety of indicators of effective teaching, relatively unrelated to a wide variety of background variables, and useful to lecturers for purposes of improving teaching effectiveness (Marsh, 1987, 1995; Marsh & Dunkin, 1992). However, there have been only limited attempts to test the applicability of North American instruments, or the generalizability of findings from North American research, in other countries. Marsh (1981) noted that there is danger in assuming that instruments developed in one setting can be used effectively in new settings without first testing their applicability. In order to address this issue, he introduced the applicability paradigm (Marsh, 1981, 1986) for studying the applicability of his Students' Evaluations of Educational Quality (SEEQ; Marsh, 1982a; 1982b, 1984, 1987; Marsh & Dunkin, 1992) instrument. The present investigation is an extension of that research based on a Chinese translation of SEEQ.

Multidimensionality of Students' Evaluations of Teaching Effectiveness

Effective teaching is a multidimensional construct Thus, it is not surprising that a considerable body of North American research has also shown that SETs are also multidimensional (see Marsh, 1987). In evaluating the need to distinguish among appropriately defined multiple dimensions, it is important to consider the purposes that the evaluations are intended to serve. Marsh (1984, 1987; also see Braskamp, Brandenburg & Ory, 1985; Centra, 1979; Doyle, 1983; McKeachie, 1979; Murray, 1980) noted that student ratings are used variously to provide: (a) formative feedback to faculty about the effectiveness of their teaching; (b) a summative measure of teaching effectiveness to be used in personnel decisions; (c) information for students to use in the selection of lecturers and courses; and (d) an outcome or a process description for research on teaching. Whereas there is some disagreement about whether a single summary score is more useful than multidimensional ratings for purposes of personnel decisions, there is

general agreement that appropriately constructed multiple dimensions are more useful for the other three purposes.

Information from SETs depends upon the content of the items. Poorly worded or inappropriate items will not provide useful information. If a survey instrument contains an ill-defined hodgepodge of different items and student ratings are summarized by an average of these items, then there is no basis for knowing what is being measured. Particularly when the purpose of the ratings is formative, it is important that careful attention be given to the components of teaching effectiveness that are to be measured. Surveys should contain separate groups of related items which are derived from a logical analysis of the content of effective teaching and the purposes which the ratings are to serve, and should be supported by empirical procedures such as factor analysis and multitrait-multimethod analysis.

The SET literature contains several examples of well constructed instruments with clearly defined factor structures that provide measures of distinct components of teaching effectiveness. In addition to his SEEQ instrument, Marsh (1987) noted Frey's Endeavor instrument (Frey, Leonard & Beatty, 1975; also see Marsh, 1981a, 1986), the Student Description of Teaching questionnaire (Hildebrand, Wilson & Dienst, 1971), and the Michigan State Student Instructor Rating System (Warrington, 1973). Factor analyses of responses to each of these instruments provided clear support for the factor structure they were designed to measure, demonstrating that the SETs do measure distinct components of teaching effectiveness. He suggested that the systematic approach used in the development of these instruments and the similarity of the factors which they measure support their construct validity.

Research Based on the SEEQ Instrument

Here we briefly summarize research based on the SEEQ instrument that is the basis of the present investigation (for more detailed summaries see Marsh, 1984, 1987, 1995; Marsh & Dunkin, 1992). In the development of SEEQ: (1) a large item pool was obtained from a literature review on instruments in current usage, and interviews with faculty and students about what they saw as effective teaching; (2) students and faculty were asked to rate the importance of items; (3) faculties were asked to judge the potential usefulness of the items as a basis for feedback; and (4) open-ended student comments were examined to determine if important aspects had been excluded. These criteria, along with psychometric

properties, were used to select items and revise subsequent versions, thus supporting the content validity of SEEQ responses (see Appendix for wording of the items).

Factor analytic support for SEEQ is particularly strong. To date, more than 30 published factor analyses of SEEQ responses have identified the factors that SEEQ is designed to measure (e.g., Marsh, 1982b; 1983, 1984; 1987; Marsh & Hocevar, 1991a; Marsh & Roche, 1993). Factor analyses of responses by 50,000 classes (representing responses to nearly 1 million SEEQ surveys) provided clear support for the SEEQ factor structure (Marsh & Hocevar, 1991a). In separate analyses of responses from 21 different groups representing different levels of instruction (e.g., undergraduate and graduate level courses) and a diversity of academic disciplines, the same set of SEEQ factors were identified. When lecturers evaluated their own teaching effectiveness on the same SEEQ form as completed by their students, factor analyses of student ratings and lecturer self-evaluations each identified the same SEEQ factors (Marsh, 1982b; Marsh, Overall & Kesler, 1979). Marsh and Bailey (1993) evaluated profiles of SEEQ responses for a cohort of 221 teachers who had been evaluated with SEEQ regularly (an average of 25 sets of ratings per teacher) over a 13 year period. Not only were ratings on separate SEEQ scales stable over time (Marsh & Hocevar, 1991b), but so were the multidimensional profiles of ratings. The profile for each teacher (e.g., high on Enthusiasm but low on Organization) was distinct from the profiles of other teachers, and generalized over time and course level. These studies demonstrate the broad generalizability of SEEQ factors over time, across academic disciplines, and across responses by students and by teachers.

In several large studies the combined effect of many potential biases explained no more than 15% of the variance in student ratings. Student ratings were: positively correlated with Prior Subject Interest, Expected Grades, and Workload/Difficulty; negatively correlated with class size. Although SEEQ research and meta-analytic reviews (Feldman, 1993, in press) have demonstrated little effect of student or teacher gender, researchers have more recently suggested that there may be a student gender by teacher gender interaction such that students give higher ratings to teachers of the same sex (Feldman, in press). Even though there are modest relations between SEEQ responses and a few background variables , a careful examination of the nature of these effects and corresponding relations with teacher self-evaluations of their own teaching suggests that they may not represent biases.

- Paradoxically, at least based upon the supposition that **Workload/Difficulty** is a potential bias, higher levels of Workload/Difficulty were positively correlated with student ratings and with teacher self-evaluations of their own teaching effectivenss.
- **Class size** is negatively correlated with student ratings of SEEQ factors most logically related to class size -- Group Interaction and Individual Rapport -- but not with other SEEQ factors. Similarly, class size is negatively correlated with teacher self-evaluations on these two factors but not other SEEQ factors. Apparently, class size has a moderate effect on these two aspects of effective teaching and these effects are accurately reflected in SETs (and teacher self-evaluations).
- For both student ratings and teacher self-evaluations, **prior subject interest** was most highly correlated with Learning/Value. Again the findings suggest that Prior Subject Interest is a variable which influences some aspects of effective teaching (particularly Learning/Value) and these effects are accurately reflected in both the SETs and teacher self-evaluations.
- Class-average **expected grades** are positively correlated with student ratings, but there are quite different explanations for this finding: (a) grading leniency hypothesis: teachers who give higher-than-deserved grades will receive higher-than-deserved student ratings -- a serious bias; (b) validity hypothesis: better Expected Grades reflect better student learning so that the effect supports the validity of SETs; and (c) student characteristics hypothesis: pre-existing student characteristics (e.g., prior subject interest) may affect student learning, student grades, and teaching effectiveness, so that the effect is explained in terms of pre-existing differences. The grade a student receives is likely to be related to the grading leniency of the teacher, how much the student learned, and characteristics that the student brings to the course. Not surprisingly there is some support for each explanation, but the clearest support is for the validity hypothesis and, perhaps, the student characteristics hypothesis. Thus grading leniency effects may produce a bias in SETs, but support for this suggestion is weak and the size of such an effect would be small.

In summary, research summarized here shows that relations with potential biasing factors tend to be small. More importantly, a more careful examination of the nature of these effects suggests that they should not be interpreted as biases.

SEEQ responses have been successfully validated in relation to learning in multisection validity studies (Marsh, Fleiner & Thomas, 1975; Marsh & Overall, 1980), the ratings of former students (Marsh, 1977; Overall & Marsh, 1980), lecturer self-evaluations of their own teaching effectiveness (Marsh, 1982b; Marsh, Overall & Kesler, 1979), affective course consequences such as plans to pursue further study (Marsh & Overall, 1980), a feedback intervention that targeted specific SEEQ scales (Marsh & Roche, 1993), and a variety of other criteria (see Marsh, 1987, for an overview). SEEQ ratings are primarily a function of the lecturer who teaches a course and not the course that is being evaluated (Marsh, 1981b; Marsh & Overall, 1981).

Feedback from SEEQ responses, particularly when coupled with a candid discussion with an external consultant, led to improved student ratings and better student learning (Marsh & Roche, 1993; Overall & Marsh, 1979; also see Cohen, 1980). In further support of the multidimensionality of SEEQ responses and their effectiveness as feedback, Marsh and Roche (1993) demonstrated that a feedback intervention that targeted specific SEEQ dimensions improved teaching effectiveness overall but had its largest effect on those specific dimensions of SEEQ that were specifically targeted.

The Applicability Paradigm

The overarching purpose of the applicability paradigm is to evaluate the appropriateness of SEEQ in new settings. In the initial applicability paradigm study Marsh (1981) asked University of Sydney students from diverse disciplines to select "one of the best" and "one of the worst" lecturers they had experienced, and to rate each on an instrument containing SEEQ and Endeavor items. As part of the study, students were asked to indicate "inappropriate" items, and to select up to five items that they "felt were most important in describing either positive or negative aspects of the overall learning experience in this instructional sequence" for each lecturer. Analyses included a discrimination analysis examining the ability of items and factors to differentiate between "best" and "worst" lecturers, a summary of "not appropriate" responses, a summary of "most important item" responses, and a MTMM analysis of agreement between responses to two different SET instruments, and factor analyses of the responses.

Marsh (1986, 1987) reviewed five applicability studies conducted in Australia, New Zealand, Spain, and Papua New Guinea. He noted that the Spanish study (Marsh, Touron, & Wheeler, 1985) was the first to use a translated version of the SET instruments and one of the Australian studies was the only

study conducted in a non-university setting, whereas Clarkson (1984) emphasized that the Papua New Guinea study was conducted in a non-Western setting. In each of the studies, all but the Workload/Difficulty items strongly differentiated between the good and poor teachers. Differences in the Workload/Difficulty items were much smaller, although the good teachers tended to teach courses that were judged to be more difficult and to have a heavier workload. SET items were judged to be "inappropriate" if a student specifically indicated the item to be inappropriate or failed to respond to the item. Results from all the studies showed that items were judged to be appropriate by 80% or more of the students, even though a few items were judged to be "inappropriate" by more than 10% of the students. (The items most frequently judged to be inappropriate came from the Group Interaction, Individual Rapport, Examination, and Assignment factors.) Students were also asked to select up to five items that were most important in describing the overall learning environment and all items were selected by at least some of the students as being most important. Across all the studies the most frequently nominated items came from the Enthusiasm, Learning/value, and Organization factors. These findings support the applicability of SEEQ in a wide range of settings.

Subsequent applications of the applicability paradigm, particularly by Watkins and colleagues, have considered responses by students from a wide variety of different countries: India (Watkins & Thomas, 1991); Nepal (Watkins & Regmi, 1992), Nigeria (Watkins & Akande, 1992), Philippines (Watkins & Gerong, 1992), and Hong Kong (Watkins, 1992). In each study, the SET items were presented in their original English version to university students who were instructed in English (even if English was not their first language). In support of the applicability and construct validity of the SEEQ responses, the studies show that: SEEQ items are broadly appropriate, SEEQ responses define factors that generalize across diverse settings, students differentiate among SEEQ factors, SEEQ factors differentiate between good and bad teachers, and that the relative importance of different SEEQ factors is similar in diverse settings. Based on his evaluation of applicability paradigm research from a cross-cultural perspective, Watkins (1994, p. 262) concluded that "the results are certainly generally encouraging regarding the range of university settings for which the questionnaires and the underlying model of teaching effectiveness investigated here may be appropriate."

Despite the generally supportive findings from this applicability research, evidence in support of the differentiation between good and poor teachers has consistently been much stronger than particularly the factor analytic evidence in support of the SEEQ factor structure and the ability of students to discriminate between different components of teaching effectiveness. However, this finding may reflect in part idiosyncrasies in the design of the applicability paradigm, the reliance on factor analyses by individual students, and methodological limitations in the factor analyses. Consistent with the design of applicability studies, it is hardly surprising that a lecturer selected as being "best" by a student is consistently rated more favorably than one who is selected as being "worst". The halo effect produced by this selection process probably exaggerates the differentiation among good and poor teachers, but also makes it more difficult to distinguish among the multiple components of effective teaching and substantially increases the size of correlations among the different factors. Hence, the differentiation is a double-edged sword; too little would suggest that the ratings are not valid, but too much would undermine support for their multidimensionality.

The distinction between the individual student and the more typical class-average unit of analysis is somewhat blurred in this research. Whereas analyses are conducted on responses by individual students, there are relatively few cases in which the different students would choose the same course when there is a sufficiently diverse sample of students. In the extreme, when there is only one student per class, the individual student and class-average response are the same. Nevertheless, larger random and systematic errors associated with individual student responses -- compared to class average responses based on groups of 20 or more students -- are likely to make the underlying factor structure more difficult to identify than the more typical factor analysis based on class-average responses. Finally, all applicability paradigms conducted thus far have relied solely on exploratory factor analysis. Whereas the advantages of recent advances of confirmatory factor analysis over exploratory factor analysis are well known (Bentler, 1990; Bollen, 1989; Byrne, 1989; Marsh, 1994), they may be particularly important in applicability studies like those summarized here.

The purpose of the present investigation is to evaluate the applicability of a Chinese translation of the SEEQ instrument at the Chinese University of Hong Kong. This study, however, differs from other applicability paradigm studies in a number of important features. First, except for the Spanish study

(Marsh, Touron, & Wheeler, 1985), this is the first of these studies to use a translated version of the SEEQ. This is particularly important in the present study since many students in this university would not be able to complete a questionnaire with full understanding when it is presented in English even though most of the students have studied English as a second language. Second, students in the present study completed SEEQ items but were not presented items from the Endeavor instrument Finally, the present investigation demonstrates important new advances in the application of CFA in evaluating the factor structure of responses to SEEQ and testing the invariance of this factor structure across responses by students from different academic disciplines (a between-group comparison) and across responses to good and poor teachers by the same student (a within-group comparison).

Method

Sample.

The sample consists of 844 (287 males and 557 females) Chinese University of Hong Kong students, representing one-third of all students in their final year of undergraduate study. Sample sizes from each of the seven faculties in the university are roughly in proportion to the number of students in that faculty: arts (143), business administration (200); education (36), engineering (85); medicine (43); science (155); and social science (182). For purposes of some analyses, the seven faculties were combined to form three discipline groups (group 1 = arts, social sciences, and education; group 2 = business administration; group 3 = engineering, medicine, and science). Whereas all 844 respondents completed the survey for a good teacher, only 825 completed the matching survey for the poor teacher. Surveys were mailed to students and returned in a postage-paid envelope. Anonymity was ensured in that students provided neither their own names nor the names of teachers who they nominated as good and poor teachers.

Materials.

Each questionnaire consisted of introductory materials that included instructions and a limited number of items requesting demographic information (faculty, class size, grade, student gender, teacher gender, and teacher age). Students were initially requested to select a good and a poor teacher based on their experience at the Chinese University of Hong Kong. Students were asked to try to limit their choices to lecturers who were in charge of an instructional sequence that lasted at least one term and who used

mainly a lecture, seminar, or discussion style of presentation. Students were then asked to complete two separate questionnaires, one each for the "good" and the "poor" teacher. The SEEQ instrument that is the focus of this study and its research basis were described earlier (see Appendix 1 for the SEEQ items and scales; for a complete copy of SEEQ and permission to use it, see Marsh & Bailey, 1993; Marsh & Roche, 1993). Students responded to items on a nine-point response scale which varied from "1--strongly disagree" to "9--strongly agree " (except for the three SEEQ items designed to measure Workload/Difficulty which have idiosyncratic responses scales -- see Appendix). An additional "not appropriate" response was provided for items judged to be not relevant to the particular course being evaluated (responses to items left blank were also counted as "not appropriate"). After completing the ratings for each teacher, students were asked to select up to five items that they felt were most important in describing either positive or negative aspects of their overall learning experience in each course.

Statistical Analysis

Preliminary analyses. Each item and SEEQ scale was tested in terms of its: (a) ability to discriminate between good and poor teachers; (b) appropriateness (i.e., a lack of "not appropriate" or missing responses); (c) importance (i.e., the number of "most important" nominations), and relation to background variables. A comparison of mean responses for good and poor teachers was conducted with paired t-tests for each SEEQ item and scale to evaluate discrimination between good and poor teachers. Simple frequencies were used to evaluate potentially not appropriate and most importance items and scales. Background variables were correlated with responses to SEEQ items and scales. Because some research (Feldman, 1993, in press) suggests that the effects of teacher and student gender may interact (e.g., students give higher ratings to same-sexed teachers), a two-way ANOVA was then used to evaluate the main and interactive effects of these variables All these preliminary statistical analyses were conducted with the Windows version of the SPSS statistical package (SPSS, 1995).

Confirmatory factor analysis. Confirmatory factor analyses (CFAs) were conducted with the SPSS version of LISREL 7 (Joreskog & Sorbom, 1989) using maximum likelihood estimates derived from covariance matrices based on pairwise deletion for missing data. A detailed description of CFA is beyond the scope of the present investigation and is available elsewhere (e.g., Bollen, 1989; Byrne, 1989; Joreskog & Sorbom, 1989; Marsh, 1994; Pedhazur & Schmelkin, 1991). Following Marsh and Balla (1994),

Marsh, Balla, and Hau (1996), and Marsh, Balla, and McDonald (1988) we emphasize the Tucker-Lewis index (TLI) to evaluate goodness of fit, but also present the chi-square test statistic, the relative noncentrality index (RNI), the root mean square error of approximation (RMSEA), the parsimony index based on the RNI (PRNI), and an evaluation of parameter estimates. Whereas there are no precise standards for what values of indices such as these are needed for an "acceptable" fit, typical guidelines are that the TLI and RNI should be greater than .9, PRNI should be greater than .8, and RMSEA should be less than .05. However, model comparison is also facilitated by positing a partially nested ordering of models in which the parameter estimates for a more restrictive model are a proper subset of those in a more general model (for further discussion see Bentler, 1990). In the present application, for example, a model in which factor loadings are constrained to be invariant across solutions for the three academic discipline groups is nested under a model in which the factor loadings are estimated freely within each of the groups. However, the fit indices for alternative models can be compared whether or not the particular models are nested. The fit indices considered here vary primarily in terms of how they incorporate parsimony. All other things being equal, more parsimonious models are preferable to more complex models. RNI does not correct for parsimony, whereas each of the other indices do. Thus a more parsimonious model may have a better index of fit based on indices that incorporate parsimony such as the TLI that is emphasized here. Whereas tests of statistical significance and indices of fit aid in the evaluation of the fit of a model, there is ultimately a degree of subjectivity and professional judgment in the selection of a "best" model.

When parallel data exist for more than one group, CFA provides a particularly powerful test of the equivalence of solutions across the multiple groups. For present purposes the invariance of factor loadings over academic discipline is a substantively important issue of particular concern to potential users of the SEEQ instrument. For tests of factorial invariance the researcher is able to fit the data subject to the constraint that any one, any set, or all parameters are equal in the multiple groups. The minimal condition for "factorial invariance" is the equivalence of all factor loadings in the multiple groups. Typically, it is also of substantive interest to test for the invariance of relations among factors (e.g., Marsh, 1994; Marsh & Hocevar, 1985). The invariance of factor variances and of the uniquenesses and correlated uniqueness associated with measured variables is typically less substantively relevant and is

likely to be idiosyncratic to particular groups (also see related discussion by Bentler, 1988; Bollen, 1989; Byrne, 1989; Byrne, Shavelson & Muthen, 1989; Joreskog & Sorbom, 1989; Marsh & Hocevar, 1985).

Although most applications of the factorial invariance consider multiple (between) group comparisons, essentially the same logic applies to within-group tests of invariance. In the present investigation, students completed SEEQ items in relation to both a good teacher and a poor teacher. Hence, it is possible to constrain factor loadings, factor correlations, factor variances, uniquenesses, and correlated uniqueness to be invariant across responses to good and poor teachers. In addition to the substantively relevant issues about factorial invariance, a methodologically interesting question concerns relations between matching items or factors for good and poor teachers. There is no a priori reason why ratings of good and poor teachers by the same student should be correlated, but it may be necessary to consider correlated uniquenesses for ratings of the same item under different conditions that reflect method/halo effects idiosyncratic to individual students.

It is also reasonable to evaluate more complex models that posit both within-and between-group invariance constraints. Thus, for example, a particularly relevant model is one in which factor loadings are constrained to be equal across all six sets of parameters representing the responses for good and poor teachers by the same students (a within-group comparison) and responses across the three different academic disciplines (a between-group comparison).

The hierarchy begins with the least restrictive model in which only the form of the model -- the number of factors and the pattern of fixed and nonfixed parameters -- is invariant across groups. This initial baseline model is "totally noninvariant" in the sense that there are no invariance constraints on estimated parameters. This model is critically important because it provides a basis of comparison for all subsequent models in the invariance hierarchy. Particularly when the focus of the research is on the invariance of parameters across groups, Marsh (1994) suggests that it may be reasonable to use theory, prior research (with different data), common sense, or -- if necessary -- empirical guidelines based on the same data (e.g., LISREL's modification indices; see Joreskog & Sorbom, 1989; 1993) to modify the a priori model and then to test the invariance constraints with this a posteriori baseline model. Because the focus of subsequent models in the hierarchy is on tests of invariance constraints imposed on the baseline model, it is useful to test the same pattern of fixed and free parameters for all groups in the a posteriori

baseline model. This also provides a test of whether the model modifications from one group generalize to another group. If, however, there are substantively important differences in the a priori and a posteriori solutions, then the a posteriori results should be interpreted cautiously and, ultimately, should be cross validated with new data. In the present investigation preliminary analyses are conducted based on the total sample to establish a baseline model and this baseline model is then evaluated for each of the three discipline groups. Consistent with recommendations by Marsh, the same baseline model (i.e., the same pattern of fixed and free parameters) is tested for each set of SEEQ responses for good and poor teachers and for each of the three discipline groups.

Results

Applicability.

The major findings about the applicability of SEEQ responses to the Chinese University of Hong Kong are summarized in Table 1. Good teachers are rated substantially higher than poor teachers on all SEEQ items and scales, but the sizes of the differences vary substantially. Thus, for example, there is relatively little difference in the ratings of good and poor teachers on Workload ratings, although good teachers tend to teach somewhat more difficult and demanding classes. The largest differences are for Enthusiasm, Learning/Value, and particularly Organization. There are also systematic differences in the variances of responses such that there is more variability in ratings of poor teachers than good teachers in both SEEQ items and scales. It is also interesting to note that correlations between responses to matching good and poor teachers selected by the same student are nearly uncorrelated for all SEEQ scales and items. The only major exception is the positive correlation ($r = .49$) for item 33 (average number of hours per week outside of class). No other correlation exceeded .20 in absolute value, most were less than .10, and about half were negative. In summary, responses to SEEQ items and scales clearly differentiated between good and poor teachers.

Insert Table 1 About Here

All SEEQ items and scales were selected by at least some students as most important. There are, however, substantial differences in the frequency with which different items and scales were nominated. As with differentiation between good and bad teachers, the most frequently nominated scales are Learning, Enthusiasm, and particularly Organization, whereas the least frequently nominated scale is

Workload/Difficulty. Although there is a general agreement in the frequency of nomination of SEEQ items and scales for good and poor teachers, Organization items are particularly likely to be nominated as important for the evaluation of poor teachers.

Most of the SEEQ items and scales were seen as appropriate by the vast majority of the students in the present investigation. There are, however, some important exceptions to this generalization. For ratings of both good and poor teachers, Exam ratings and particularly item 25 (feedback on exams/graded materials were valuable) were seen as not appropriate by nearly one-quarter of the students. The apparent explanation is that a number of courses only have a final examination that typically is not returned to students, whereas some students selected a teacher that only taught a component of a course for which there may have been no examinations at all. It is also relevant to note that only 5.3% (for poor teachers) and 3.1% (for good teachers) indicated that all the items in this scale were inappropriate. For good teachers, no other items were seen to be inappropriate by more than 10% of the students and most percentages are much smaller. For poor teachers, however, there were some items from both the Individual Rapport and Breadth of Coverage factors that were seen to be inappropriate by 10% or more of the students. Whereas there are generally similar values for good and poor teachers, items from the Individual Rapport factor are judged to be inappropriate far more frequently for the poor teachers than for good teachers. This suggests that circumstances (e.g., part-time teachers) are such that particularly accessibility outside of class was not seen as appropriate for teachers selected as poor teachers. Although students indicated these items as inappropriate more frequently when rating poor teachers, it may also be one reason why these teachers were selected as being poor at teaching.

Also included in Table 1 are values for some of the background variables. Good teachers, compared to poor teachers, tended to teach slightly smaller classes, to be slightly younger, and to be somewhat more likely to be female, but these differences were all very small. Students received higher grades from good teachers than from poor teachers, and this relation was larger than those based on other background variables considered here. To the extent that the course grade is a reflection of course mastery and achievement, this result supports the validity of the ratings. If, however, the higher grades reflect easier grading standards, then this relation may reflect a bias in the ratings.

Insert Table 2 About Here

The correlations between these background variables and student ratings of good and poor teachers (Table 2) are also informative. Grades tend to be positively correlated with all the SEEQ scales, but the size and even direction varies depending on the scale. Grades are most highly correlated with the overall course rating and the Learning factor, and this is what might be expected if expected grades reflect mastery and achievement. Also, for both good and poor teachers, grades are substantially more highly correlated with Learning and Overall Course ratings than with teacher Enthusiasm and Overall Teacher ratings. This also suggests that the grade effect supports the validity of student ratings rather than a teacher grading leniency effect. It is also interesting to note that, like North American research (e.g., Marsh, 1987), grades are negatively correlated with Workload/Difficulty -- courses in which students earn poorer grades are seen as more difficult. The pattern of relations with expected grades and, perhaps, interpretations of these results for Chinese University students is similar to North American SEEQ research.

Class size (Table 2) is also modestly correlated with ratings of good teachers, though the sizes of the correlations are small and several are not statistically significant. As might be expected, teachers with larger class sizes are rated somewhat lower in terms of Group Interaction and Individual Rapport. For poor teachers, however, the correlations with class size are smaller and one of the two significant correlations is positive. The only significantly negative correlation with class size for poor teachers is for Group Interaction. This general pattern of results for class size for Chinese University students is also consistent with findings from North American research using SEEQ that was discussed earlier.

The remaining background variables in Table 2 have to do with student gender, teacher gender, and their interaction. Whereas most of these effects are not statistically significant, there are slight tendencies for female students to give higher ratings for both good and poor teachers, and for women teachers to receive higher ratings. However, the critical interaction effects are all small and 20 of 22 fail to reach statistical significance. For the two interaction effects that are statistically significant, male students give higher ratings to female teachers than male teachers, whereas female students give slightly higher ratings to male teachers than female teachers Hence there is little or no evidence that students provide higher ratings to same-sexed teachers. These results, consistent with North American research,

suggest that gender of the student and the teacher have little effect on student ratings by Chinese University students.

SEEQ Factor Structure.

Baseline model. In the initial baseline model the two sets of 35 SEEQ items were posited to define 18 SEEQ factors -- 9 for good teachers and 9 for poor teachers. In this model, each SEEQ item (except for the overall ratings) was allowed to load on only the one factor that it was intended to measure, correlations among the 18 factors were freely estimated, and the uniquenesses associated with each item were uncorrelated with other uniquenesses. Whereas the goodness of fit of this very restrictive model is reasonable (TLI = .871, Model 1a, Table 3), inspection of LISREL's modification indices indicated that the inclusion of additional correlated uniquenesses would improve the fit of the model. Typically, when participants respond to the same instrument under two different conditions (for good and poor teachers in the present investigation) there are correlated uniquenesses associated with responses to the same item in different conditions. Surprisingly, the only substantial modification index for matching items from the good and poor teacher responses was item 33 (average number of hours per week outside of class; also see earlier discussion of Table 1). However, the modification indices were substantial for several pairs of items within responses for good teachers or within responses for poor teachers. In each case, these results indicated that two items within the same SEEQ scale were more highly correlated than could be explained in terms of the mutual reliance on the underlying common factor that each was designed to measure (see Table 4 for a listing of the particular correlated uniquenesses that were included). In keeping with the decision to maintain a common baseline model for each set of SEEQ responses, parameters freed for responses to either the ratings of good or poor teachers were freed for both sets of items. The goodness of fit for this final baseline model (TLI = .921, Model 1b Table 3) was substantially improved and adequate by traditional guidelines (TLI > .90). Models 1b, 1c and 1d differed from the final baseline model by excluding respectively the correlated uniquenesses for good teacher ratings, for poor teacher ratings, or the one correlated uniqueness associated with the same item in the good and poor ratings. In each case, the fit was poorer, thus supporting the inclusion of all these correlated uniquenesses in the final baseline model.

Insert Tables 3 and 4 About Here

In order to conserve space, the solution for this final baseline model is not presented because the parameter estimates are substantially similar to those presented in Table 4 (which are based on subsequent analyses of the same data for each of the three discipline groups). The factor solution (not shown) is fully proper (e.g., there are no Heywood cases), factor loadings are all statistically significant and substantial (e.g., a majority of the factor loadings for items designed to measure specific factors are greater than .7 in standardized form). Although all 9 SEEQ factors are positively correlated for both good teacher factors and poor teacher factors, none of these factor correlations exceeds .80 and most are less than .50. Hence, the results of this final baseline model provides clear support for the nine SEEQ factors. Interestingly, correlations between the good and poor teacher factors are close to zero.

Tests of invariance. All analyses considered in this section were conducted on a set of three covariance matrices, one for each of the three discipline groups discussed earlier. In the least restrictive of these models, the final baseline model was fit to each discipline group with no constraints such that parameter estimates are the same across any of the groups. The fit of this "no invariance" model (TLI=.919, Model 2a in Table 3) is good and provides an important basis of comparison for more restrictive models that impose invariance constraints. We also report the fit of the final baseline model for each group separately (models 2b, 2c, and 2d in Table 3). Because the TLIs are each greater than .9 for each discipline group, the results support the fit of the baseline model for each of the three groups. We now turn to invariance tests in which parameter estimates are required to be the same in different groups, a major focus of this study.

In models 3a - 3e, various sets of parameters are required to be equal across the three discipline groups (the between-group invariance constraints in Table 3). In the first and, perhaps, most critical between-group invariance model, the factor loadings were constrained to be the same across the three groups. The TLI for this model is the same as the model with no invariance constraints, thus providing good support for the invariance of the factor loadings. In the next three between group invariance models, factor variances, factor correlations, and both factor correlations and factor variances are constrained to be equal across the three discipline groups. For each of these models, the TLI is marginally higher than for the no invariance model and is best for the model constraining both factor correlations and variances to be equal across groups (TLI = .921, model 3d). The further imposition of invariances of uniquenesses

and correlated uniquenesses still provided an acceptable fit (TLI = .910), but one that was lower than the no-invariance comparison model as well as model 3d. In summary, these results provide clear support for the invariance of factor loadings, factor correlations, and factor variances across the three discipline groups, but not, perhaps, the invariance of the uniquenesses.

In models 4a - 4e, various set of parameters were required to be equal across the matching parameter estimates for ratings of good and poor teachers. Because each student made ratings of both good and poor teachers, these are referred to as within-group constraints in Table 3. Although the same type of constraints are considered as for the between-group constraints, support for the invariance of the parameter estimates is not so strong for the within-group constraints. Although there may be reasonable support for the invariance of the factor loadings and factor correlations, there is no support for the invariance of the factor variances or the uniquenesses. This result is consistent with earlier observations that responses to SEEQ items and scales had much more variability for ratings of poor teachers than good teachers.

In models 5a - 5h, between-group invariance constraints over the different discipline groups (as in models 3a - 3e) and within-group constraints over responses to good and poor teachers (as in models 4a - 4e) were considered simultaneously. Although it would be possible to test a very large number of different models, we have focused primarily on those constraints that were supported in the earlier analyses. Comparison of the TLIs for the alternative models provides support for factor loadings, factor variances, and factor correlations across the three discipline groups and across at least the factor loadings for ratings of good and poor teachers (TLI = .918, model 5d in Table 3). Whereas it may be reasonable to further argue for the invariance of factor correlations across ratings of good and poor teachers (TLI = .917), it is evident that the imposition of further between- or within-group invariance constraints leads to a noticeable decrement in the TLIs. Hence, the interpretation of these models imposing both between- and within-group invariance constraints is consistent with models evaluating either between- or within-group constraints separately.

For purposes of illustration, we have chosen to present the factor solution in which factor loadings, factor correlations, and factor variances are constrained to be invariant across the discipline groups and only factor loadings constrained to be invariant across ratings of good and poor teachers.

Because of the nature of the between group invariance constraints, it is only necessary to present one set of factor loadings, factor variances, and factor correlations (because they are the same across the three discipline groups), but three sets of uniquenesses and correlated uniquenesses. As already noted, inspection of these parameter estimates provide strong support for the SEEQ factor structure for ratings of both good and poor teachers in that all the factor loadings are statistically significant and a majority are greater than .70. Also, although there are a few substantial factor correlations, none of the correlations are greater than .8 and most are less than .5.

The comparison of the parameter estimate across ratings of good and poor teachers is somewhat complicated by differences in the metric of the ratings and the particular strategy used by LISREL to standardize the ratings. When there are multiple groups, LISREL's completely standardized parameter estimates are standardized in relation to the pooled variance estimate across all three groups so that parameters from the multiple groups are directly comparable. This common metric is important when evaluating invariance constraints. The squared multiple correlations (SMRs in Table 4) are the square of what the factor loadings would be if each item was standardized in relation to responses in just its own group rather than the pooled responses.

For the within-group comparisons, however, the ratings of good and poor teachers are standardized in relation to only their own idiosyncratic variances so that the completely standardized parameter estimates do not vary along a common metric. This is particularly evident for the comparison of factor loadings for the good-teacher factors and the poor teacher factors. Even though these factor loadings were constrained to be invariant in the original, common metric, there are small but systematic differences in the completely standardized factor loadings that are a function of the standardization procedure. Factor correlations, however, are comparable across the ratings by good and poor teachers. Although the correlations among good teacher factors differ from those among poor teacher factors, these differences do not appear to be substantial. This observation is consistent with the finding that the added imposition of the invariance of factor correlations across factors for good and poor teachers did not substantially influence the goodness of fit (TLIs of .917 vs. .918). It is also relevant to compare the factor variances for the good and poor teachers. Consistent with earlier observations based on raw item and scale scores (Table 1), there is substantially more variability in the poor teacher factors. In relation

to factor variances for good teacher factors that were fixed at 1.0, the factor variances for the poor teacher factors are typically twice as large.

Discussion

The purposes of this study were to evaluate the applicability of a Chinese translation of SEEQ, to explore the generality of findings based on North American research to a Hong Kong setting, and to demonstrate recent advances in the application of CFA to the study of university students' evaluations of teaching effectiveness. The results provided good support for the Chinese SEEQ. Good teachers were consistently rated much more favorably than poor teachers on all SEEQ items and scales, demonstrating that SEEQ responses do differentiate between good and poor teachers as defined here. Whereas most of the SEEQ items were judged to be appropriate by most of the Chinese University students, there was at least one notable exception. Items from the Examination scale -- particularly the item on feedback from examinations -- was judged to be inappropriate by approximately one-quarter of the students. This suggests that responses to this scale will have to be interpreted cautiously and, perhaps, disregarded if there are a high proportion of missing values in a particular class. All SEEQ items and scales were nominated as most important by at least some Chinese University students, but items from the Learning, Enthusiasm, and particularly the Organization scales were nominated most frequently. Because this general pattern of results has been reported consistently in the applicability paradigm studies reviewed earlier, the present results support the generality of previous research to this new setting.

The relations between SEEQ responses and background variables (class size, expected grade, Workload, student gender, and teacher gender) and their interpretation -- particularly in relation to potential biases -- is relevant to evaluating the generalizability of findings from the large volume of research conducted primarily in North America. Workload tends to be positively related to all SEEQ scales, which is opposite to what would be predicted by a bias hypothesis. Class size tends to be negatively correlated with SEEQ responses, but only the relations with Group Interaction and Individual Rapport were large enough to warrant attention. Grades were modestly correlated with SEEQ responses, but the pattern of results suggests that this relationship may support the validity of the ratings rather than a bias in the ratings due to the teacher's grading leniency. Finally, there was no support at all for the hypothesis that male and female students give higher ratings to teachers of the same gender. Because

these results are also consistent with generalizations based on North American research, these results again support the generality of previous research findings in this new setting.

Although CFA and related statistical techniques are widely applied in social science research and Marsh (1987, 1991a, 1991b) noted the need to apply this technique more widely in SET research, there has been surprisingly little such research. In particular, CFA has apparently not been applied in any of the previous applicability paradigm studies that were one basis of the present investigation. The many advantages of CFA are likely to be particularly important for applicability studies. Asking students to select a good and a poor teacher who are then evaluated reinforces halo effects such that the good teacher is evaluated favorably on all items whereas the poor teacher is evaluated negatively on all items. Hence, it may not be surprising that the exploratory factor analyses that have been used in this research are sometimes unable to identify all the a priori factors. Furthermore, exploratory factor analysis is inherently weak for purposes of comparing the factor structure over responses to good and poor teachers or comparing the factor structure in different academic disciplines. Hence, despite the fact that the present study was one of the very few applicability paradigms to be conducted in a non-Western setting and one of the very few where the SEEQ was translated into a language other than English, the results provide perhaps the strongest support for the a priori SEEQ structure and its generalizability over academic disciplines. Many of the critical questions about the invariance of the SEEQ factor structure over different groups of teachers that were a major focus of the present investigation could not be addressed adequately with exploratory factor analyses like those presented in previous applicability paradigm studies. These results are also of practical significance in the Chinese University of Hong Kong where many faculties felt that different instruments would have to be constructed for each faculty. (Actually SEEQ offers a compromise in that there is also provision for faculties, departments, or individual teachers to select items from an item bank or construct their own items to supplement the common core of SEEQ items considered here). In this sense, the present investigation offers a potentially useful demonstration of the usefulness of CFA and its application.

Author Note

The authors would like to thank Lawrence Roche and Alexander S. Yeung for their helpful comments on earlier versions of this article. Correspondence concerning this article should be addressed to Professor Herbert W. Marsh, Faculty of Education, University of Western Sydney, Macarthur, PO Box 555, Campbelltown, NSW 2560 Australia.

References

Bentler, P. M. (1988). Theory and implementation of EQS. A structural equations program. Los Angles: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indices in structural models. Psychological Bulletin, 107, 238-246.

Bollen, K. A. (1989). Structural equations with latent variables. New York: John Wiley & Sons.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial invariance. Psychological Bulletin, 105, 456-466.

Byrne, B. M. (1989). A primer of LISREL: Basic applications and programming for confirmatory factor analytic models. New York: Springer Verlag.

Braskamp, L. A., Brandenburg, D. C. & Ory, J. C. (1985). Evaluating teaching effectiveness: A practical guide. Beverly Hills, CA: Sage.

Centra, J. A. (1979). Determining faculty effectiveness. San Francisco, Jossey-Bass.

Clarkson, P. C. (1984). Papua New Guinea Students' Perceptions of Mathematics lecturers. Journal of Educational Psychology, 76, 1386-1395.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis. Research in Higher Education, 13, 321-341.

Doyle, K. O. (1983). Evaluating teaching. Lexington, MA: Lexington Books.

Feldman, K. A. (1993). College students' views of male and female college teachers. Part II -- evidence from students' evalutions of their classroom teachers. Research in Higher Education, 34, 151-211.

Feldman, K. A. (in press). Reflections on the study of effective college teaching and student ratings: One continuing quest and two unresolved issues. In J. Smart (Ed.), Higher education: Handbook of theory and research (Vol. 13, pp. xxx-xxx). New York: Agathon.

Frey, P. W., Leonard, D. W., & Beatty, W. W. (1975). Student ratings of instruction: Validation research. American Educational Research Journal, 12, 327-336.

Hildebrand, M., Wilson, R. C. & Dienst, E. R. (1971). Evaluating university teaching. Berkeley: Center for Research and Development in Higher Education, University of California, Berkeley.

Joreskog, K. G. (1971). Statistical analyses of sets of congeneric tests. Psychometrika, 36, 109-134.

Joreskog, K. G., & Sorbom, D. (1989). LISREL 7: A guide to the program and applications. Chicago:SPSS, Inc.

Joreskog, K. G., & Sorbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Chicago: Scientific Software International.

Marsh, H. W.(1977). The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. American Educational Research Journal, 14, 441-447.

Marsh, H. W.(1981). Students' evaluations of tertiary instruction: Testing the applicability of American surveys in an Australian setting. Australian Journal of Education, 25, 177-192.

Marsh, H. W. (1982a). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. British Journal of Educational Psychology 52, 77-95.

Marsh, H. W. (1982b). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. Journal of Educational Psychology, 74, 264-279.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. Journal of Educational Psychology, 75, 150-166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76, 707-754.

Marsh, H. W. (1986). Applicability paradigm: Students' evaluations of teaching effectiveness in different countries. Journal of Educational Psychology, 78, 465-473.

Marsh, H. W.(1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11, 253-388. (Whole Issue No. 3)

Marsh, H. W. (1991a). A multidimensional perspective on students' evaluations of teaching effectiveness: A reply to Abrami and d'Apollonia (1991). Journal of Educational Psychology, 83, 416-421.

Marsh, H. W. (1991b). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. Journal of Educational Psychology, 83, 285-296.

Marsh, H. W. (1993). Relations between global and specific domains of self: The importance of individual importance, certainty, and ideals. Journal of Personality and Social Psychology, 65, 975-992.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. Structural Equation Modeling, 1, 5-34.

Marsh, H. W. (1995). Student evaluation of teaching. In T. H. Husen & T. N. Postlethwaite (Eds.), International encyclopedia of education. Oxford: Pergamon Press.

Marsh, H. W., & Bailey, M. (1993). Multidimensionality of students' evaluations of teaching effectiveness: A profile analysis. Journal of Higher Education, 64, 1-18.

Marsh, H. W., & Balla, J. (1994). Goodness of fit in confirmatory factor analysis: The effects of sample size and model parsimony. Quality and Quantity: International Journal of Methodology, 28, 185-217.

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E. Schumacker (Eds.), Advanced structural equation modeling techniques (pp. 315-353). Hillsdale, NJ: Erlbaum.

Marsh, H. W., Balla, J., & McDonald, R. P. (1988). Goodness of fit in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. Smart (Ed.), Higher education: Handbook of theory and research (Vol. 8 pp. 143-233). New York: Agathon.

Marsh, H. W., Fleiner, H. & Thomas, C. S. (1975). Validity and usefulness of student evaluations of instructional quality. Journal of Educational Psychology, 67, 833-839.

Marsh, H. W., & Hocevar, D. , D. (1985). The application of confirmatory factor analysis to the study of self-concept: First and higher order factor structures and their invariance across age groups. Psychological Bulletin, 97, 562-582.

Marsh, H. W., & Hocevar, D. (1991a). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. Teaching and Teacher Education, 7, 9-18.

Marsh, H. W., & Hocevar, D. (1991b). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. Teaching and Teacher Education, 7, 303-314.

Marsh, H. W., & Overall, J. U. (1980). Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. Journal of Educational Psychology, 72, 468-475.

Marsh, H. W., & Overall, J. U. (1981). The relative influence of course level, course type, and instructor on students' evaluations of college teaching. American Educational Research Journal, 18, 103-112.

Marsh, H. W., Overall, J. U. & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. Journal of Educational Psychology, 71, 149-160.

Marsh, H. W., & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. American Educational Research Journal, 30, 217-251.

Marsh, H. W., Touron, J. & Wheeler, B. (1985). Students' evaluations of university instructors: The applicability of American instruments in a Spanish setting. Teaching and Teacher Education, An International Journal of Research and Studies, 1, 123-138.

McKeachie, W.J. (1979). Student ratings of faculty: A reprise. Academe, 384-397.

Murray, H. G. (1980). Evaluating university teaching: A review of research. Toronto, Canada: Ontario Confederation of University Faculty Associations.

Overall, J. U. & Marsh, H. W., (1979). Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. Journal of Educational Psychology, 71, 856-865.

Overall, J. U. & Marsh, H. W., (1980). Students' evaluations of instruction: A longitudinal study of their stability. Journal of Educational Psychology, 72, 321-325.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

SPSS (1991). SPSS user's guide. Chicago: SPSS, Inc.

Warrington, W. G. (1973). Student evaluation of instruction at Michigan State University. In A. L. Sockloff (Ed.), Proceedings: The first invitational conference on faculty effectiveness as evaluated by students (pp. 164-182). Philadelphia: Measurement and Research Center, Temple University.

Watkins, D. (1992). Evaluating the effectiveness of tertiary teaching: A Hong Kong perspective. Educational Research Journal, 7, 60-67.

Watkins, D. (1994). Student evaluations of teaching effectiveness: A Cross-cultural perspective. Research in Higher Education, 35, 251-266.

Watkins, D. & Akande, A. (1992). Student evaluations of teaching effectiveness: A Nigerian investigation. Higher Education, 24, 453-463.

Watkins, D. & Gerong, A. (1992). Evaluating tertiary teaching: A Filipino investigation. Educational and Psychological Measurement, 52, 727-734.

Watkins, D. & Regmi, M. (1992). Student evaluation of tertiary teaching: A Nepalese investigation. Educational Psychology, 12, 131-142.

Watkins, D. & Thomas, B. (1991). Assessing teaching effectiveness: An Indian perspective. Assessment and Evaluation in Higher Education, 16, 185-198.

Table 1

Discrimination Between Good and Poor Teachers, Most Important Items, Not Appropriate Items

	good		poor				% Judged To Be: Most Impt		Not Approp	
	Mn	SD	Mn	SD	r	t-value	Good	Poor	Good	Poor
Lrn	6.93	1.00	3.82	1.42	.03	52.12	57.3	50.7	0.0	0.0
1	6.71	1.44	4.36	2.02	.02	27.51	17.4	9.6	1.2	0.8
2	7.30	1.27	4.12	1.99	.01	39.09	32.2	22.8	0.0	0.1
3	6.89	1.38	2.98	1.63	.01	52.46	18.9	23.6	0.4	0.4
4	6.82	1.22	3.83	1.86	.10	40.50	8.9	13.1	0.1	0.4
Enth	7.40	.98	3.18	1.44	-.09	67.04	62.7	60.4	0.0	0.2
5	7.88	1.05	3.89	2.02	-.03	49.50	34.0	25.1	0.0	0.6
6	7.75	1.13	3.43	1.87	.00	56.66	18.5	12.4	0.0	0.4
7	6.98	1.58	3.00	1.80	-.12	45.04	20.1	9.5	0.2	0.6
8	7.00	1.25	2.41	1.39	-.11	66.40	8.4	30.9	0.9	0.4
Org	7.34	.94	3.08	1.41	-.05	70.34	58.3	70.6	0.0	0.1
9	7.56	1.02	2.82	1.64	-.05	68.76	31.6	43.1	0.0	0.5
10	7.51	1.09	2.95	1.68	-.01	64.83	26.0	29.6	0.0	0.5
11	7.28	1.20	3.39	1.78	-.05	50.56	7.2	10.2	0.5	0.7
12	6.99	1.48	3.13	1.94	-.03	43.86	9.0	21.4	3.2	2.9
Grp	6.70	1.40	3.54	1.79	.03	40.03	26.6	21.4	1.2	2.3
13	6.56	1.66	3.64	2.11	.06	31.43	10.9	6.7	3.9	5.2
14	6.52	1.69	3.61	2.02	.07	31.45	5.7	4.5	5.1	6.2
15	6.89	1.44	3.35	1.88	.02	41.78	6.5	9.0	3.0	4.4
16	6.86	1.53	3.56	1.94	-.02	37.02	8.3	4.7	3.1	5.2
Ind	7.27	1.16	4.14	1.66	-.02	43.77	45.6	29.4	0.0	1.0
17	7.64	1.29	4.49	2.06	.01	36.85	22.9	15.3	0.2	2.2
18	7.77	1.19	4.44	1.92	.02	39.81	18.5	5.4	2.0	11.0
19	6.82	1.51	3.36	1.76	.00	40.27	10.9	10.8	2.8	10.3
20	6.91	1.64	4.18	1.97	-.03	26.40	7.4	5.8	5.3	20.6
Brd	6.89	1.07	3.74	1.61	.00	45.99	22.0	18.1	1.2	3.9
21	6.96	1.23	3.74	1.81	-.01	39.13	7.1	7.0	6.6	8.7
22	7.06	1.22	3.74	1.84	.01	41.46	7.8	4.7	5.0	6.7
23	6.83	1.37	3.84	1.85	.06	35.67	5.1	3.2	8.6	12.5
24	6.76	1.43	3.64	1.96	-.02	34.03	6.7	6.3	7.0	10.9
Exam	6.74	1.23	4.09	1.71	.03	36.10	17.7	26.5	3.1	5.3
25	6.65	1.52	3.35	1.77	.04	33.10	4.4	10.4	24.3	27.4
26	7.04	1.28	4.02	1.97	.07	34.91	8.8	15.4	10.3	14.1
27	6.66	1.71	4.53	2.07	-.07	21.57	6.1	7.1	4.7	7.6

Table 1 (continued)

	good		poor				% Judged To Be: Most Impt		Not Approp	
	Mn	SD	Mn	SD	r	t-value	Good	Poor	Good	Poor
Asgn	6.95	1.23	4.46	1.89	.02	31.16	17.4	14.0	2.8	5.1
28	6.95	1.30	4.55	1.96	.04	28.70	10.2	9.6	3.9	5.8
29	6.97	1.31	4.38	1.96	.01	30.48	8.8	6.7	3.8	6.4
Work	5.08	1.02	4.58	1.21	.14	9.82	2.2	3.0	0.2	0.8
30	5.72	1.36	5.40	1.65	-.08	3.99	.9	1.3	1.1	1.7
31	5.49	1.57	4.80	1.65	.04	8.73	1.1	.5	3.1	4.0
32	5.28	.89	4.90	1.63	.10	5.84	.6	1.2	0.6	3.9
33	3.84	2.07	3.17	2.11	.49	9.97	.1	.7	2.8	10.8
Crs 34	7.24	1.19	2.97	1.46	-.15	59.93	.1	.7	1.1	0.8
Tch 35	7.83	.93	2.58	1.32	-.18	86.30	1.5	2.8	0.4	1.0
Background Variables										
Size	62.10	42.96	66.17	42.88	.45	-2.55				
Grade	7.75	1.39	6.00	2.07	.26	-23.05				
Age	41.03	8.18	42.80	7.77	.16	-4.95				
Gender	1.17	.38	1.09	.28	.14	5.63				

Note. Lrn = Learning, Enth = Enthusiam, Org = Organization, Grp = Group Interaction, Ind = Individual Rapport, Brd = Breadth, Exam =Examinations, Asgn = Assignments, Work = Workload/Difficulty, Crse = Overall Course Rating, Tch = Overall Teacher Rating. Results are presented for a scale score and the items comprising each SEEQ factor (see Appendix). Minimum Ns are 844 for Good teachers and 824 for poor teachers. The % missing values for each item and scale are presented as "not appropriate" responses. Results for a paired t-test was used to compare responses to good and poor teachers are summarized as the r (correlation between responses to good and poor teachers by the same student) and the t-value (all t-values significant at $p < .01$). Percentages of cases indicating an item to be "most important" or "not appropriate" are also presented for all but the background variables (Size = class size, Age = teacher age, Gender = teacher gender (1=male, 2=female). For each scale score the % for most important responses is the % selecting at least one item from that scale and the % for not appropriate is the % indicating not appropriate for all items from that scale.

Table 2

Correlations Between Student Ratings and Background Variables

	Correlations					Beta Weights		
	Student Gender	Class Size	Grade	Teacher Age	Teacher Gender	Student Gender	Teacher Gender	Interaction
Good								
Learning	.08*	-.11**	.27**	.02	.02	.07	.03	-.07
Enthusiasm	.08*	-.06	.04	-.00	-.06	.08*	-.05	-.05
Organization	.10**	-.04	.12**	.00	.04	.08*	.04	-.06
Group Interact	.18**	-.17**	.03	-.06	.11**	.16**	.11*	-.06
Individ Rapport	.11**	-.18**	.12**	-.05	.02	.09**	.03	-.08*
Breadth	.07*	-.08*	.10**	.01	-.02	.08*	-.04	.01
Exams	.04	-.13**	.19**	-.02	-.00	.04	-.01	.01
Assignments	.07*	-.07	.11**	.02	.07*	.06	.07	-.02
Workload	.06	.03	-.08*	.10**	.09**	.05	.07	.03
Overall Course	.01	-.11**	.35**	.07*	-.07	.03	-.07	.01
Overall Teacher	.01	-.09**	.14**	.02	-.07*	.02	-.07	-.01
Poor								
Learning	.03	.05	.16**	-.07*	.13**	.02	.13*	-.01
Enthusiasm	.09*	-.02	.02	-.01	.05	.08*	.04*	-.02
Organization	.10**	-.02	.08*	-.05	.11**	.09*	.09*	.02
Group Interact	.09**	-.07*	.05	-.10**	.19**	.07*	.17*	.01
Individ Rapport	.06	-.02	.08*	.02	.09**	.04	.12*	-.09*
Breadth	.08*	.02	.05	-.00	.06	.07	.07	-.05
Exams	.05	.01	.15**	.03	.07	.05	.04	.05
Assignments	.05	-.04	.13**	.03	.04	.04	.05	-.03
Workload	.06	.07	-.21**	-.10**	.04	-.06	.04	.02
Overall Course	.05	.08*	.18**	-.09**	.09**	.04	.10*	-.02
Overall Teacher	.10**	-.02	.07*	.00	.06	.08*	.07	-.05

Note. Background variables are scaled such that positive correlations reflect higher ratings associated with female students, larger class sizes, higher grades, older teachers, and female teachers. Beta weights are standardized beta weights when student ratings are predicted from student gender, teacher gender, and their interaction.

* $p < .05$; ** $p < .01$.

Table 3

Goodness of Fit For Between- and Within-Group Invariance Constraints on a Common Baseline Model

Model	χ^2	df	RNI	TLI	RMSEA	PRatio	PRNI	Model Description/Invariance Constraints
Total Group Baseline Model								
	20585.56	2415	.000	.000	.123	1.000	.000	Null Total Group
1a	4300.46	2184	.884	.871	.044	.904	.799	Initial Baseline Total Group
1b	3472.74	2177	.929	.921	.035	.901	.837	Final Baseline Total Group
1c	3747.28	2182	.914	.905	.038	.904	.826	M2a with no Correlate Unique (CUs) for Good Teachers
1d	3914.10	2182	.905	.894	.040	.904	.817	M2a with no CUs for Bad Teachers
1e	3628.32	2178	.920	.911	.037	.902	.830	M2a with no CUs between Good and Bad Teachers
Baseline Model For Each Group - No Invariance								
	24689.77	7245	.000	.000	.070	1.000	.000	Null Model, sum across 3 groups
	11225.67	2415	.000	.000	.127	1.000	.000	Null Model - group 1
	6114.54	2415	.000	.000	.118	1.000	.000	Null Model - group 2
	7349.56	2415	.000	.000	.114	1.000	.000	Null Model - group 3
2a	7797.98	6531	.927	.919	.020	.901	.836	Baseline Model, sum across 3 groups
2b	2945.62	2177	.913	.903	.040	.901	.823	Baseline Model - group 1
2c	2403.22	2177	.939	.932	.031	.901	.846	Baseline Model - group 2
2d	2449.14	2177	.945	.939	.028	.901	.852	Baseline Model - group 3
Between Group (BG) Invariance								
3a	7946.16	6643	.925	.919	.020	.917	.848	Factor Loading (FL) Inv
3b	8283.83	6949	.923	.920	.020	.959	.886	FL, Factor Corr (FC) Inv
3c	7964.44	6679	.926	.920	.020	.922	.854	FL, Factor Variances (FV) Inv
3d	8315.80	6985	.924	.921	.020	.964	.891	FL, FC, FV Inv
3e	8702.21	7147	.911	.910	.021	.986	.899	Total Between Group Inv
Within Group (WG) Invariance								
4a	7942.92	6615	.924	.917	.020	.913	.844	FL Inv
4b	8058.03	6723	.923	.918	.020	.928	.857	FL, FC Inv
4c	8149.97	6642	.914	.906	.021	.917	.838	FL, FV Inv
4d	8408.59	6750	.905	.898	.022	.932	.843	FL, FC, FV Inv
4e	10151.18	7104	.825	.822	.029	.981	.809	Total Within Group Inv
BG and WG Invariance								
5a	8369.04	6977	.920	.917	.020	.963	.886	BG Inv: FL, FR; WG Inv: FL
5b	8419.78	7013	.919	.917	.020	.968	.890	BG Inv: FL, FR; WG Inv: FL, FR
5c	8764.98	7040	.901	.898	.022	.972	.876	BG Inv: FL, FR; WG Inv: FL, FR, FV
5d	8404.24	7013	.920	.918	.020	.968	.891	BG Inv: FL, FR, FV; WG Inv: FL
5e	8453.11	7049	.920	.917	.020	.973	.895	BG Inv: FL, FR, FV; WG Inv: FL, FR
5f	8602.35	7022	.909	.907	.021	.969	.881	BG Inv: FL, FR, FV; WG Inv: FL, FV
5g	8786.96	7058	.901	.898	.022	.974	.878	BG Inv: FL, FR, FV; WG Inv: FL, FR, FV
5h	10495.20	7260	.815	.815	.030	1.002	.816	BG Inv: Total; WG Inv: Total

Note. df = degrees of freedom, RNI = Relative Noncentrality Index, TLI = Tucker-Lewis Index, RMSEA = Root Mean Square Error of Approximation, PRatio = Parsimony ratio (df model/df null model), PRNI = Parsimony Index based on RNI (PRatio x RNI). Initial baseline models (M1a - M1e) were based on analyses of the total group covariance matrix but all other models are based on the common baseline model fit to separate covariance constraints for each group with alternative between group or within group invariance constraints.

Table 4

Factor Solution For Model 4d (see Model 5d in Table 3)

		Good Teacher factors									Poor Teacher Factors									Group 1			Group 2			Group 3		
		Lrn	Enth	Org	Grp	Ind	Brd	Exm	Asgn	Wrk	Lrn	Enth	Org	Grp	Ind	Brd	Exm	Asgn	Wrk	SMR	Uniq	CU[a]	SMR	Uniq	CU[a]	SMR	Uniq	CU[a]
Good Teacher Ratings																												
Lrn	1	.65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.43	.56		.44	.55		.40	.63	
	2	.81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.68	.31		.70	.27		.60	.45	
	3	.68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.50	.47		.42	.63		.46	.56	
	4	.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.25	.74		.30	.57		.22	.90	
Enth	5	0	.68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.48	.51		.49	.48		.44	.60	
	6	0	.67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.44	.50	.16	.49	.46	.16	.44	.57	.30
	7	0	.45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.22	.72		.18	.91		.19	.84	
	8	0	.58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.36	.59	.25	.30	.76	.46	.32	.71	.33
Org	9	0	0	.71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.47	.57		.56	.40		.52	.46	
	10	0	0	.80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.64	.36		.63	.37		.63	.38	
	11	0	0	.74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.58	.41		.56	.43		.51	.53	
	12	0	0	.56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.32	.68		.34	.60		.29	.75	
Grp	13	0	0	0	.78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.65	.34		.59	.43		.58	.45	
	14	0	0	0	.78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.65	.33	.13	.63	.35	.15	.54	.51	.18
	15	0	0	0	.87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.82	.17		.77	.22		.68	.36	
	16	0	0	0	.90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.83	.16		.76	.25		.79	.22	
Ind	17	0	0	0	0	.78	0	0	0	0	0	0	0	0	0	0	0	0	0	.62	.37		.73	.22		.53	.54	
	18	0	0	0	0	.86	0	0	0	0	0	0	0	0	0	0	0	0	0	.74	.26		.79	.20		.71	.31	
	19	0	0	0	0	.69	0	0	0	0	0	0	0	0	0	0	0	0	0	.46	.56		.50	.47		.48	.52	
	20	0	0	0	0	.61	0	0	0	0	0	0	0	0	0	0	0	0	0	.35	.71		.53	.33		.34	.71	
Brd	21	0	0	0	0	0	.67	0	0	0	0	0	0	0	0	0	0	0	0	.44	.57		.45	.54		.45	.54	
	22	0	0	0	0	0	.75	0	0	0	0	0	0	0	0	0	0	0	0	.58	.41	.21	.55	.47	.05	.56	.45	.16
	23	0	0	0	0	0	.77	0	0	0	0	0	0	0	0	0	0	0	0	.62	.37		.61	.38		.56	.47	
	24	0	0	0	0	0	.67	0	0	0	0	0	0	0	0	0	0	0	0	.47	.49		.47	.49		.39	.70	
Exm	25	0	0	0	0	0	0	.74	0	0	0	0	0	0	0	0	0	0	0	.62	.33		.57	.41		.46	.65	
	26	0	0	0	0	0	0	.79	0	0	0	0	0	0	0	0	0	0	0	.64	.35		.67	.30		.57	.46	
	27	0	0	0	0	0	0	.45	0	0	0	0	0	0	0	0	0	0	0	.19	.87		.21	.75		.22	.73	
Asgn	28	0	0	0	0	0	0	0	.86	0	0	0	0	0	0	0	0	0	0	.82	.16		.63	.44		.74	.26	
	29	0	0	0	0	0	0	0	.90	0	0	0	0	0	0	0	0	0	0	.83	.16		.75	.26		.80	.20	
Wrk	30	0	0	0	0	0	0	0	0	.60	0	0	0	0	0	0	0	0	0	.37	.60		.39	.55		.32	.77	
	31	0	0	0	0	0	0	0	0	.72	0	0	0	0	0	0	0	0	0	.56	.42		.59	.36		.46	.63	
	32	0	0	0	0	0	0	0	0	.42	0	0	0	0	0	0	0	0	0	.17	.90		.20	.70		.18	.80	
	33	0	0	0	0	0	0	0	0	.37	0	0	0	0	0	0	0	0	0	.14	.82		.18	.65		.12	.07	
Crs	34	.40	.18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.30	.65		.33	.57		.23	.94	
Tch	35	0	.53	0	0	.15	0	0	0	0	0	0	0	0	0	0	0	0	0	.41	.58	.27	.47	.46	.06	.35	.73	.32

Table 4 (continued)

		Good Teacher factors									Poor Teacher Factors									Group 1			Group 2			Group 3		
		Lrn	Enth	Org	Grp	Ind	Brd	Exm	Asgn	Wrk	Lrn	Enth	Org	Grp	Ind	Brd	Exm	Asgn	Wrk	SMR	Uniq	CU[a]	SMR	Uniq	CU[a]	SMR	Uniq	CU[a]
Poor Teacher Ratings																												
Lrn	1	0	0	0	0	0	0	0	0	0	.64	0	0	0	0	0	0	0	0	.42	.55		.39	.64		.40	.61	
	2	0	0	0	0	0	0	0	0	0	.77	0	0	0	0	0	0	0	0	.60	.39		.58	.41		.57	.45	
	3	0	0	0	0	0	0	0	0	0	.77	0	0	0	0	0	0	0	0	.60	.39		.67	.30		.54	.50	
	4	0	0	0	0	0	0	0	0	0	.47	0	0	0	0	0	0	0	0	.21	.87		.24	.72		.25	.68	
Enth	5	0	0	0	0	0	0	0	0	0	0	.60	0	0	0	0	0	0	0	.33	.72		.44	.45		.35	.67	
	6	0	0	0	0	0	0	0	0	0	0	.65	0	0	0	0	0	0	0	.38	.70	.40	.50	.42	.24	.45	.52	.33
	7	0	0	0	0	0	0	0	0	0	0	.60	0	0	0	0	0	0	0	.32	.78		.38	.59		.42	.49	
	8	0	0	0	0	0	0	0	0	0	0	.73	0	0	0	0	0	0	0	.48	.57	.36	.58	.39	.21	.58	.38	.13
Org	9	0	0	0	0	0	0	0	0	0	0	0	.69	0	0	0	0	0	0	.43	.64		.53	.42		.52	.44	
	10	0	0	0	0	0	0	0	0	0	0	0	.80	0	0	0	0	0	0	.60	.42		.67	.32		.67	.32	
	11	0	0	0	0	0	0	0	0	0	0	0	.77	0	0	0	0	0	0	.60	.39		.63	.35		.55	.48	
	12	0	0	0	0	0	0	0	0	0	0	0	.65	0	0	0	0	0	0	.41	.60		.43	.56		.44	.55	
Grp	13	0	0	0	0	0	0	0	0	0	0	0	0	.80	0	0	0	0	0	.63	.37		.68	.30		.62	.39	
	14	0	0	0	0	0	0	0	0	0	0	0	0	.81	0	0	0	0	0	.61	.42	.17	.74	.23	.12	.65	.34	.17
	15	0	0	0	0	0	0	0	0	0	0	0	0	.86	0	0	0	0	0	.75	.25		.70	.31		.76	.24	
	16	0	0	0	0	0	0	0	0	0	0	0	0	.93	0	0	0	0	0	.88	.12		.86	.14		.87	.13	
Ind	17	0	0	0	0	0	0	0	0	0	0	0	0	0	.75	0	0	0	0	.51	.55		.59	.39		.64	.32	
	18	0	0	0	0	0	0	0	0	0	0	0	0	0	.89	0	0	0	0	.81	.18		.76	.24		.76	.25	
	19	0	0	0	0	0	0	0	0	0	0	0	0	0	.86	0	0	0	0	.73	.27		.70	.31		.77	.22	
	20	0	0	0	0	0	0	0	0	0	0	0	0	0	.76	0	0	0	0	.58	.42		.57	.42		.56	.45	
Brd	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.74	0	0	0	.53	.49		.62	.33		.52	.50	
	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.81	0	0	0	.65	.35	.23	.69	.29	.16	.63	.39	.27
	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.87	0	0	0	.74	.26		.81	.17		.73	.28	
	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.76	0	0	0	.59	.40		.48	.62		.64	.32	
Exm	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.85	0	0	.72	.28		.69	.33		.76	.23	
	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.75	0	0	.61	.36		.49	.59		.56	.44	
	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.56	0	0	.31	.69		.29	.76		.33	.64	
Asgn	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.93	0	.90	.10		.89	.10		.80	.22	
	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.92	0	.87	.13		.81	.20		.85	.15	
Work	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.53	.28	.74		.28	.72		.30	.68	
	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.75	.63	.34		.45	.69		.59	.39	
	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.27	.06	1.08		.08	.80		.08	.80	
	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.40	.19	.72	.35	.19	.69	.35	.13	1.10	.68
Crs	34	0	0	0	0	0	0	0	0	0	.45	.21	0	0	0	0	0	0	0	.43	.47		.34	.69		.29	.87	
Tch	35	0	0	0	0	0	0	0	0	0	0	.58	0	0	.17	0	0	0	0	.47	.55	.22	.58	.34	.14	.44	.60	.16

Factor Correlations

	Lrn	Enth	Org	Grp	Ind	Brd	Exm	Asgn	Wrk	Lrn	Enth	Org	Grp	Ind	Brd	Exm	Asgn	Wrk
Lrn	1																	
Enth	.63	1																
Org	.45	.74	1															
Grp	.37	.58	.37	1														
Ind	.21	.61	.42	.53	1													
Brd	.46	.55	.56	.49	.37	1												
Exm	.39	.55	.55	.48	.58	.56	1											
Asgn	.44	.39	.47	.37	.33	.44	.50	1										
Wrk	.18	.09	.07	.13	.10	.10	.11	.08	1									
Lrn	.03	.06	.09	.07	.07	.06	.17	.07	-.06	1								
Enth	-.08	-.20	-.07	-.00	-.03	-.02	.06	.01	-.06	.58	1							
Org	-.01	-.09	-.06	.03	.01	-.01	.06	.04	-.05	.52	.76	1						
Grp	-.05	-.12	-.02	-.02	-.02	-.02	.08	.01	-.08	.33	.67	.46	1					
Ind	.06	.01	.01	.05	-.01	.10	.04	.06	-.10	.27	.58	.45	.51	1				
Brd	-.05	-.09	-.05	.02	.04	-.00	.08	.00	-.10	.41	.73	.67	.57	.47	1			
Exm	.02	-.06	.06	0	.04	.01	.11	.04	.01	.39	.53	.50	.37	.48	.52	1		
Asgn	.04	.04	.11	.05	.01	.05	.14	.03	-.03	.51	.31	.39	.26	.30	.36	.54	1	
Wrk	.11	.17	.13	.11	.06	.02	.10	.10	.03	.16	.07	.10	.16	.11	.09	.12	.07	1

Factor Variances

Fac Vars	1	1	1	1	1	1	1	1	1	2.01	2.36	2.37	1.77	2.47	2.57	2.06	2.51	1.18

Note. Lrn=Learning, Enth=Enthusiasm, Org=Organization, Grp = Group Interaction, Ind = Individual Rapport, Brd = Breadth, Exm =Examinations, Asgn = Assignments, Wrk =Workload/Difficulty, Crse = Overall Course Rating, Tch = Overall Teacher Rating, SMR = squared multiple R (proportion of true score variance in each item), Uniq = uniqueness, CU = correlated uniqueness. Completely standardized parameter estimates are presented for the model with factor loadings, factor correlations, and factor variances invariant across the three discipline groups and factor loadings invariant across the solutions for good and poor teachers (see Model 5d in Table 3). Because uniquenesses and uniqueness covariances are allowed to vary across groups, separate estimates are presented for each group. The squared multiple correlation for each item is the proportion of true score variance.

[a] For ratings of good and poor teachers correlated uniquenesses are estimated between uniquenesses associated with 5 pairs of items (6 & 5; 8 & 7; 22 & 21; 35 & 34; 14 & 13) for good and for poor teachers, and for responses to item 33 for good and poor teachers.

Appendix

The SEEQ Items (paraphrased) and Scales

Learning/Value

1 Course challenging & stimulating
2 Learned something valuable
3 Increase subject interest
4 Learned & understood subject matter

Instructor Enthusiasm

5 Enthusiastic about teaching
6 Dynamic and energetic
7 Enhanced presentation with humor
8 Teaching style held your interest

Organization/Clarity

9 Teacher explanations clear
10 Materials well explained & prepared
11 Course objectives stated & pursued
12 Lectures facilitated taking notes

Group Interaction

13 Encouraged class discussion
14 Students shared knowledge/ideas
15 Encouraged questions & gave answers
16 Encouraged expression of ideas

Individual Rapport

17 Friendly towards individual students
18 Welcomed students seeking help/advice
19 Interested in individual students
20 Accessible to individual students

Breadth of coverage

21 Contrasted various implications
22 Gave background of ideas/concepts
23 Gave different points of view
24 Discussed current developments

Examinations/Grading

25 Examination feedback valuable
26 Evaluation methods fair/appropriate
27 Tested course content as emphasised

Assignments/Readings

28 Readings/texts were valuable
29 They contributed to understanding

Workload/Difficulty

30 Course difficulty (easy-hard)
31 Course workload (light-heavy)
32 Course pace (slow-fast)
33 Hours per week outside of class

Overall Rating Items

34 Overall Course Rating
35 Overall Teacher Rating

UNIVERSITY OF WESTERN SYDNEY

HERB MARSH
BA(Hons)Indiana, M Psych, PhD UCLA, DSc UWS, FASSA
Professor of Education
Dean, Graduate Research Studies

P.O. Box 555, Campbelltown, NSW 2560, Australia.
Telephone: +61 2 9772 9633. Fax: +61 2 9774 2390.
e-mail: h.marsh@uws.edu.au.

ERIC Clearinghouse on Tests, Measurement & Evaluation
American Inst. for Research
3333 K St., NW
Washington, DC 20007

I. DOCUMENT IDENTIFICATION:

Title: Students' Evaluations of University Teaching	
Author(s): Marsh, H. W.	
Corporate Source:	Publication Date: 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

Check here
Permitting microfiche (4" x 6" film), paper copy, electronic, and optical media reproduction

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here
Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Herbert W Marsh	Position: Professor of Education
Printed Name: Herbert W Marsh	Organization: Univ of Western Sydney
Address: Faculty of Educ, Univ of Western Sydney, Macarthur, PO Box 555 Campbelltown NSW 2560 Australia	Telephone Number: (02) 772 9229
	Date: 29 Jan 1997